Deployment

Pablo Bolta, Jacinto Dobón & Jorge López. Mayo 2023

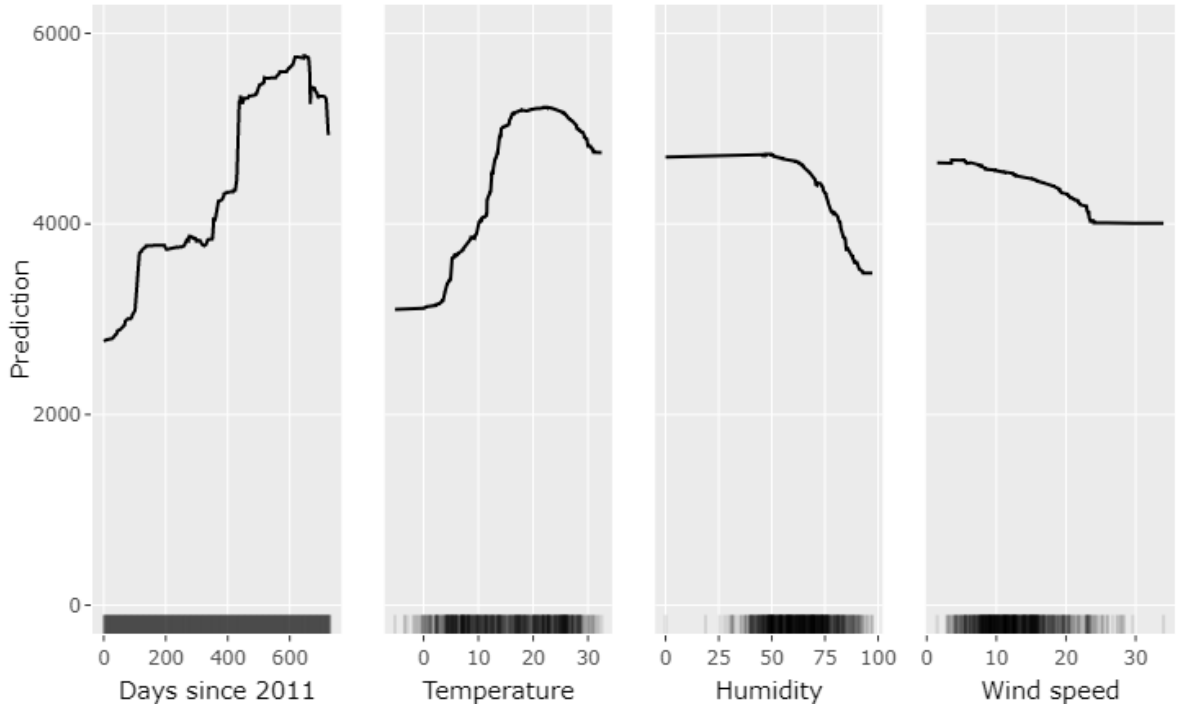
1.- One dimensional Partial Dependence Plot.

The partial dependence plot shows the marginal effect of a feature on the predicted outcome of a previously fit model.

EXERCISE:

Apply PDP to the regression example of predicting bike rentals. Fit a random forest approximation for the prediction of bike rentals (cnt). Use the partial dependence plot to visualize the relationships the model learned. Use the slides shown in class as a model.

First, we loaded the data from the data folder using DVC. The data folder contains both .csv files and is stored in our Google Drive repository. After loading the data, we transformed the variables and filtered by date from 2011 onwards. Once we finished the transformation and filtering, we trained the model using Random Forest. We then used the model to make predictions. Finally, we plotted the results using the partial dependence plot for four variables: "days_since_2011", "temp", "hum", and "windspeed".



QUESTION:

Analyse the influence of days since 2011, temperature, humidity and wind speed on the predicted bike counts.

'Days since 2011': *As we move forward in time (since 2011), bicycle rentals tend to increase. It is important to highlight two fundamental aspects:*

1. *Between 130 and 350 days, the difference in importance is minimal in this variable (always around 3700-3800 more bicycles).*
2. *The trend is always increasing, except in the final stretch. From day 648 onwards, the number of bicycles rented decreases from 5679 to 5104.*

These explanations are reliable, since we have observations for all values of this variable.

'Temperature': *The most remarkable aspect of this variable is that as the temperature increases, more bicycles are rented, from 3180 bicycles at -5 degrees to 5274 bicycles at 20 degrees. In general, more bicycles are rented when the temperature is pleasant (16-26 degrees). When it is cold or hot, fewer and fewer bicycles are rented.*

'Humidity': *For humidity values below 50%, the number of bicycles rented remains constant (about 4700 bicycles). On the other hand, as humidity increases (from 50%), fewer and fewer bicycles are rented, reaching its minimum (3704 bicycles) when humidity reaches 97%. However, these explanations should be taken with caution when humidity is below 37% or above 92%, as in those cases there are not many observations, making the explanations not entirely reliable.*

'Wind speed': *The model predicts that the trend of this variable is clearly decreasing. As wind speed increases, fewer bicycles are rented, going from 4636 with wind speed equal to 1.5; to 4178 when wind speed has a value of 24. For values greater than 24, the number of bicycles rented remains stable at 4178. However, for values greater than 24, the reliability of these explanations must be reconsidered, as there are very few observations.*

2.- Bidimensional Partial Dependency Plot.

EXERCISE:

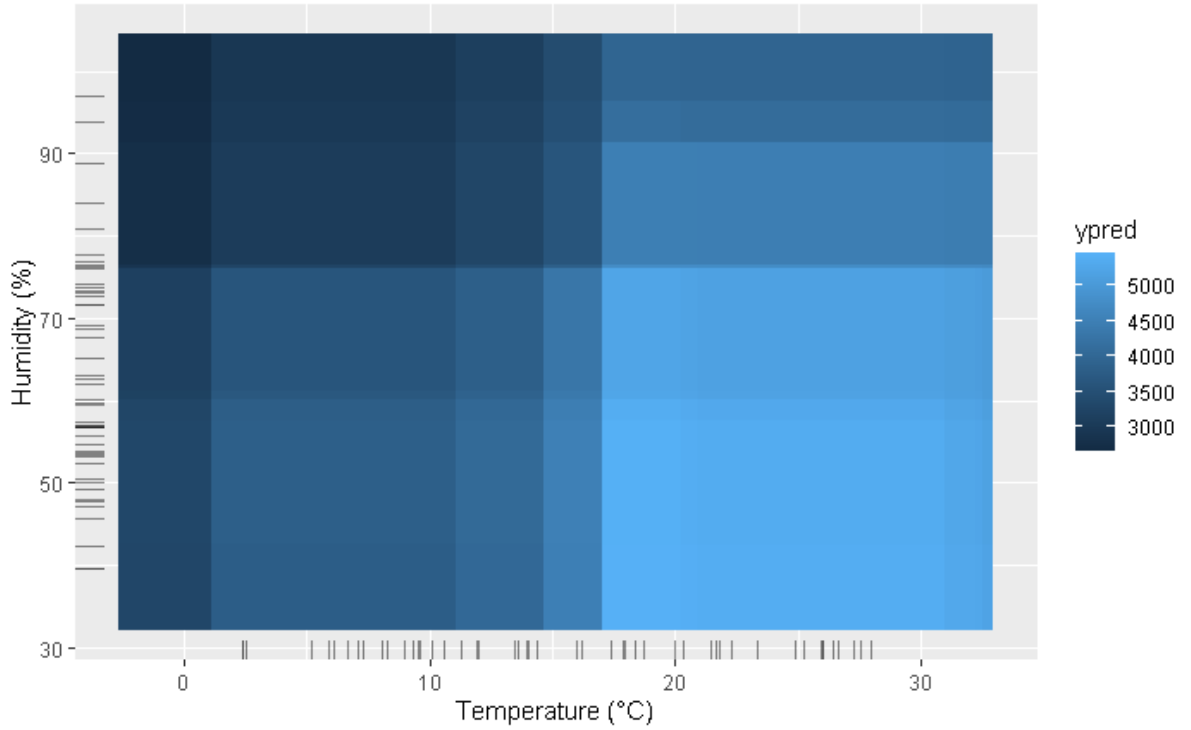
Generate a 2D Partial Dependency Plot with humidity and temperature to predict the number of bikes rented depending on those parameters.

BE CAREFUL: due to the size, extract a set of random samples from the BBDD before generating the data for the Partial Dependency Plot.

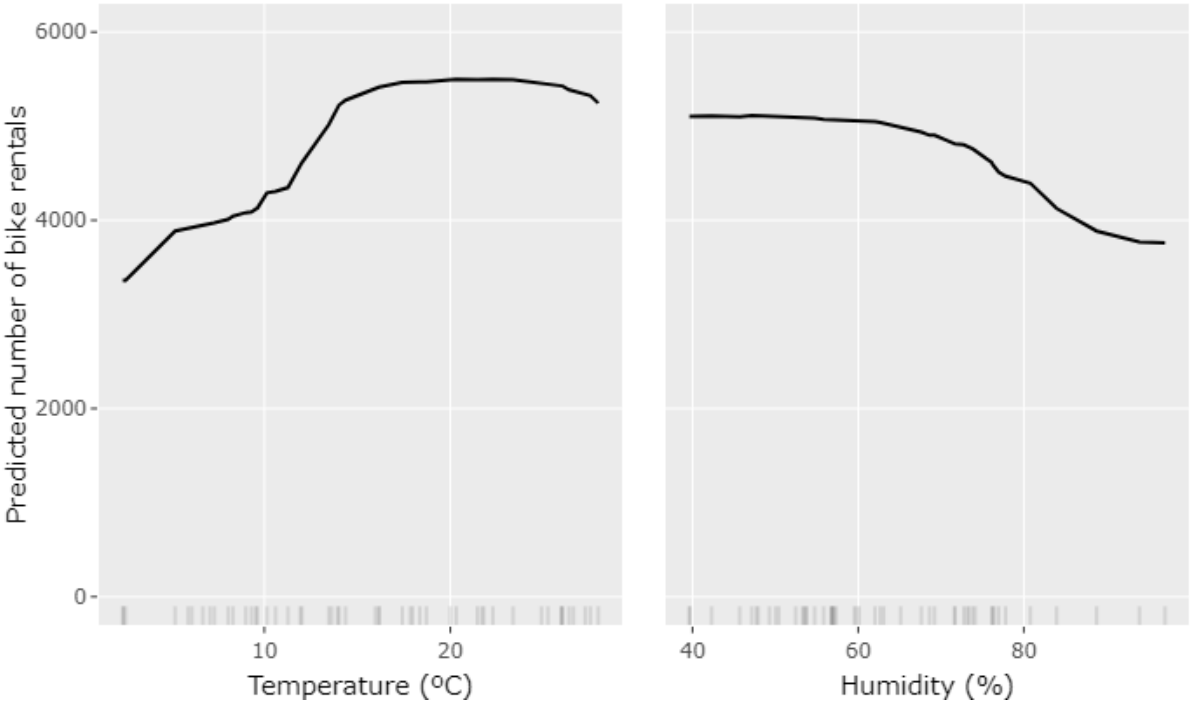
Show the density distribution of both input features with the 2D plot as shown in the class slides.

TIP: Use geom_tile() to generate the 2D plot. Set width and height to avoid holes.

We will utilize the random forest model that we previously trained to predict temperature and humidity data. To manage the large amount of data, we randomly extracted 50 observations and completed the dataset to represent the 2D Partial Dependency Plot. Our results were plotted using the two-dimensional partial dependence plot for the "temp" and "hum" variables.



Finally, we decided to include a one-dimensional plot of the extracted samples to ensure that the shape of the data is preserved.



QUESTION:

Interpret the results.

In the 2D PDP, we observe the same phenomena as in the PDP of temperature and humidity. The maximum number of bike rentals occurs at a temperature of around 20 degrees Celsius and a humidity below 50%. On the other hand, the minimum number of bike rentals occurs when the temperature is extremely low and the humidity is very high. This confirms what we saw in the PDPs, where we observed a certain independence between the effect of temperature and humidity.

While the humidity is below 50%, the maximum number of bike rentals is achieved. However, the number of rentals decreases as the humidity increases beyond this threshold. In contrast, starting from the minimum temperature, an increase in temperature leads to an increase in bike rentals until it reaches a maximum at around 17.5 degrees Celsius. Beyond this point, the number of bike rentals remains constant until the temperature reaches 25 degrees Celsius, after which it decreases again.

Based on the observations from the 1-dimensional PDPs and the 2D PDP The effects of temperature and humidity are independent. However, it is important to note that the model has not been trained on certain scenarios where there are not many real observations, and therefore, the explanations may not be reliable. It is also important to know the number of observations for each possible interaction of the variables in the dataset. Therefore, to make more accurate predictions, it may be necessary to use or collect more data, or refine the model to include more scenarios.

3.- PDP to explain the price of a house.

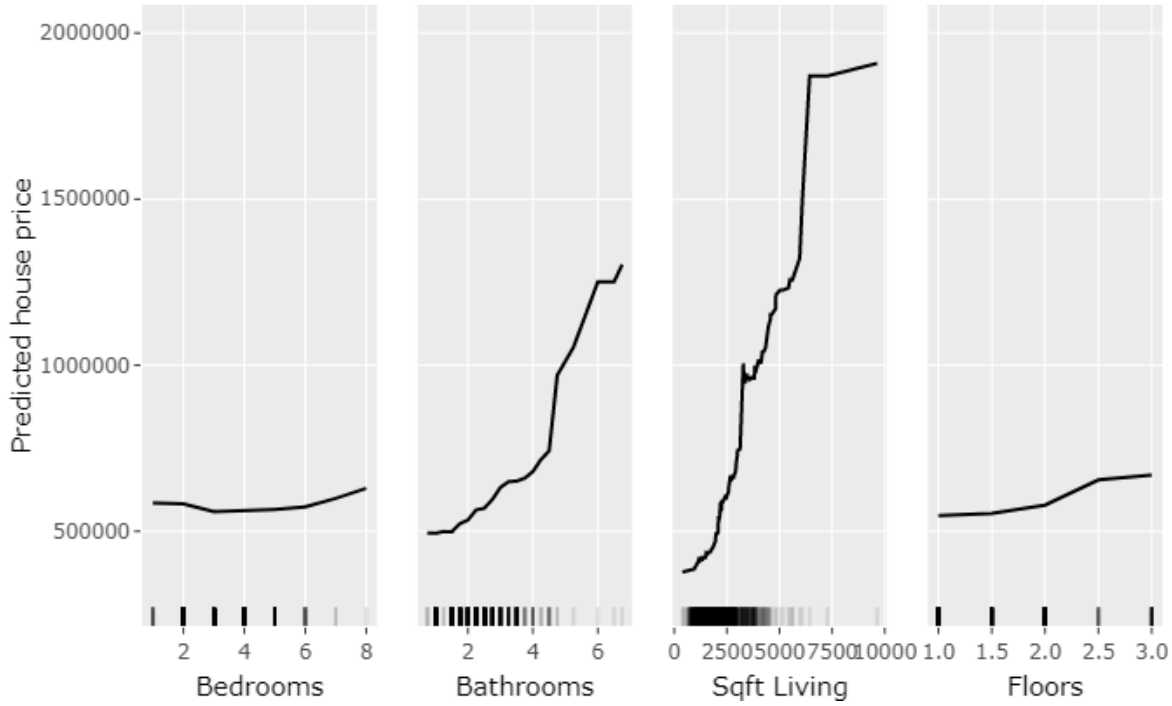
EXERCISE:

Apply the previous concepts to predict the price of a house from the database kc_house_data.csv. In this case, use again a random forest approximation for the prediction based on the features bedrooms, bathrooms, sqft_living, sqft_lot, floors and yr_built.Use the partial dependence plot to visualize the relationships the model learned.

BE CAREFUL: due to the size, extract a set of random samples from the BBDD before generating the data for the Partial Dependency Plot.

First we load the data from the dataset, extracting 1000 random samples to reduce the size. Next we obtain the dataframe that we will use for the representation.

We plot the PDPs of the proposed variables as we did in exercise 1, but with a sampled version to consume less system resources.



QUESTION:

Analyse the influence of bedrooms, bathrooms, sqft_living and floors on the predicted price.

'bedrooms': *This variable behaves in an interesting way, starting from 1 bedroom, the more bedrooms, the cheaper the house, up to 3 bedrooms. From this point, the trend reverses, and the more bedrooms, the higher the price. This is probably because 3 bedrooms is the most common option. Although it cannot be stated with total certainty, since there are few observations for the variable when there are more than 6 bedrooms.*

'bathrooms': *It is clearly observed that the more bathrooms, the higher the housing price. However, this statement is not entirely reliable in the case of the outliers 0.75 and 1.25, or for more than 4 bathrooms, since there are hardly any training samples with these values.*

'Sqft_living': *We clearly see that the larger the living area, the more expensive the housing. However, this statement may not be applicable to values below 52 square meters and above 450 square meters, which if they occur in the full data set are probably outliers. The local minimum around 33 square meters is striking, which again could be due to the fact that it is frequent in the housing stock.*

'Floors': *In this variable, we can assert that the higher the number of floors, the higher the price of the property, and this is a reliable explanation, as there are observations for all values in the training set.*